

DOCUMENT RESUME

ED 461 265

FL 024 545

AUTHOR Mislevy, Robert J.
TITLE Test Theory and Language Learning Assessment.
PUB DATE 1994-09-00
NOTE 29p.; In: Language Aptitude Invitational Symposium Program Proceedings (Arlington, VA, September 25-27, 1994); see FL 024 538. Figures may not reproduce well. Also published in "Language Testing," 12(3), p341-369, 1995.
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Evaluation Criteria; Inferences; Information Needs; *Language Aptitude; Language Proficiency; *Language Tests; Measurement Techniques; Models; Statistical Analysis; *Test Theory; *Testing; Weighted Scores

ABSTRACT

Recent developments in cognitive and educational psychology, such as increased appreciation of the situated nature of learning and understanding, call for broader ranges of student models and types of data than those standard in testing today. We must specify how what we observe on the test is related to competence as we conceptualize it, and construct a framework for carrying out inferences. There is a growing need for testing that verifies performance within a given situation or context, or in the completion of a complex task requiring the language learner to integrate several aspects of learning. A broader view of testing to include these issues requires clear definition of what is to be measured, new methods of eliciting evidence, and methods of weighting various sources of evidence for decision-making about test results. For the latter, it is possible to analyze evidence statistically so the mathematical models mirror the inference model based on that evidence. A conceptualization of test theory designed to address issues of weight and coverage of evidence for statements framed in more recent educational/psychological paradigms is outlined, with examples. (Contains 26 references.) (Author/MSE)

TEST THEORY AND LANGUAGE LEARNING ASSESSMENT

by
Robert J. Mislevy
Educational Testing Service

FL 0044545

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY

Robert J. Mislevy

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

BEST COPY AVAILABLE

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as
received from the person or organization
originating it

☐ Minor changes have been made to improve
reproduction quality

• Points of view or opinions stated in this docu-
ment do not necessarily represent official
OERI position or policy

TEST THEORY AND LANGUAGE LEARNING ASSESSMENT¹

Robert J. Mislevy
Educational Testing Service

HOLMES: *In solving a problem of this sort, the grand thing is to be able to reason backward. That is a very useful accomplishment but people do not use it much. In everyday affairs of life it is more useful to reason forward, and so the other comes to be neglected. There are fifty who can reason synthetically for one who can reason analytically.*

WATSON: *I confess I do not follow you.*

HOLMES: *I hardly expected that you would. Let me see if I can make it clearer. Most people, if you describe a train of events to them will tell you what the results would be. They can put those events together in their mind, and argue from them that something will come to pass. There are few, however, who, if you told them a result, would be able to evolve from their own inner consciousness what the steps were which led up to that result. This power is what I mean when I talk about reasoning backward, or analytically.*

WATSON: *I understand. (Doyle, 1930, p. 268).*

Introduction

Test theory, as we usually think of it, is part of a package. It encompasses models and methods for drawing inferences about what students know and can do—as cast in a particular framework of ideas from measurement, education, and psychology that coalesced in the first third of the twentieth century. In a nutshell, (i) human abilities were viewed as traits, or ‘relatively stable characteristics of a person—attributes, enduring processes, or dispositions—which are consistently manifested to some degree when relevant, despite considerable variation in the range of settings and circumstances’ (Messick, 1989, p. 15), (ii) traits were conceived as numbers along measurement scales, locating people along continua of mental characteristics just as their heights and weights located them along continua of physical characteristics, (iii) tendencies in behavior in samples of a domain of discrete settings and circumstances (e.g., assessment tasks) were the privileged form of evidence about traits, and (iv) the purpose of test theory was to guide reasoning from observed behavior in samples of situations from the domain to inferences about traits.

This ‘domain-behavior’ framework of assessment generates a universe of discourse: the nature of the problems one perceives, the kinds of statements one makes about students, the ways one gathers data to support them. Test theory, as we usually think of it, is the

¹ In addition to being presented as a plenary address at the CALL 1994 Language Aptitude Invitational Symposium, this paper also appeared in *Language Testing* 12 (3), pp. 341-369. I am grateful for discussions with Nancy Anderson, Dan Eignor, Anne Harvey, and Ming Mae Wang.

application of inferential principles to deal with such problems as missing data, source unreliability, multi-stage inference, conflicting or overlapping observations, multiple sources of disparate evidence, and constrained resources for gathering and evaluating information—as they arise in this framework.

The views of the nature and the acquisition of competence in a second language, and the nature of inferences we would wish to make about students' developing competence, do not always fall within this familiar realm. In particular, we may wish to take into account the situated and contextual aspects of language learning, and we may wish to gather data from complex tasks that stress the inter-connections among aspects of students' competence. But while these developments may suggest student models and observational strategies quite different from those employed by Spearman, Thurstone, and Thorndike, practical work under alternative perspectives inevitably faces in some form the same general inferential problems listed above. This is where a more broadly construed conception of test theory is required. It is not sufficient merely to define the class of conjectures about student competence we wish to address, and devise settings in which students can display these competencies. We must, further, specify how what we observe is related to competence as we choose to conceive it, and construct a framework for carrying out inference within the framework we thus erect.

To this end, the following section discusses the notions of evidence and inference more broadly than they are usually conceived in educational assessment. The role of probability-based inference in assessment is described. Ideas are then illustrated with two language-learning assessment challenges—contextual effects on learning and complex performance tasks—with regard to inference in the conceptual framework of the American Council of Teachers of Foreign Languages (ACTFL) guidelines (ACTFL, 1989).

Evidence and Inference

Inference is reasoning from what we know and what we observe to explanations, conclusions, or predictions. The skills we must apply in educational assessment are essentially the same as those employed in such fields as troubleshooting, medical diagnosis, criminology, and intelligence analysis. We attempt to establish the weight and coverage of evidence in what we observe. The very first question we must address is 'Evidence about what?' Schum (1987, p. 16) points out the crucial distinction between *data* and *evidence*. 'A datum becomes evidence in some analytic problem when its *relevance* to one or more hypotheses being considered is established. ...[E]vidence is relevant on some hypothesis if it either increases or decreases the likeliness of the hypothesis. Without hypotheses, the relevance of no datum could be established.'

Test data acquire meaning only in relation to particular hypotheses, or conjectures, that we entertain. The same observation can be direct evidence for some conjectures and indirect evidence for others, and wholly irrelevant to still others. In educational assessment, we construct our conjectures around notions about the nature and the acquisition of competence. We can actually observe only the specific actions and products that students

produce in specific circumstances. To evaluate their progress or guide further instruction, however, we talk at a higher level of abstraction, using specific observations as evidence for our inferences.

A conception of competence is effected as a set of variables in a student model, a simplified description of selected aspects of the infinite varieties of skills and knowledge that characterize real students. Depending on our purposes, we might distinguish anywhere from one or hundreds of facets. They might be expressed in terms of numbers, categories, or some mixture; they might be conceived as persisting over long periods of time, or apt to change at the next problem-step. They might concern tendencies in behavior, conceptions of phenomena, available strategies, or levels of development. The point is that we don't observe these variables directly. We observe only student's behavior in limited circumstances—indirect evidence about competence more abstractly conceived. Test theory, broadly construed, is conceptual and statistical machinery for reasoning from observations to inferences in terms of the competence model.

Suppose we want to make a statement about Jasmine's proficiency, in terms of likely values of the variables in a model built around some key aspects of competence. We can't observe these values directly,² but perhaps we can make an observation that bears information about the plausibility of various values under the model: her answer to a multiple-choice question, say, or two sets of judges' ratings of her violin solo, or an essay outlining how to determine which paper towel is most absorbent. The observation can't tell us her value with certainty, because similar behavior could be produced by students with different underlying levels of competency depending on factors such as their familiarity with the context and situation. It is, however, more likely to be produced by students at some levels than others. Nonsensically answering '¿Como está usted?' with 'Me llamo Carlos,' for example, is much more likely from a student classified as a Low-Novice under the American Council of Teachers of Foreign Languages (ACTFL) Reading guidelines (see Table 1) than an Advanced student. (There are similarly conceived guidelines for Writing, Speaking, and Listening.)

Conceptions of Competence

A conception of student competence and a purpose for assessment should drive the particular methods we need to get students to act in ways that reveal something about their competencies, or the forms of assessment we employ. This section contrasts key aspects of two broadly cast assessment paradigms, which we shall refer to as the 'domain-behavior' and 'cognitive/developmental,' paradigms, and notes some implications for assessment forms and test theory.

² After all, the model itself isn't truth but a simplified approximation we have constructed, and variable values are not so much characteristics of Jasmine, but of summaries of our knowledge about patterns we perceive in Jasmine's behavior, as seen through the lens of the model.

The 'domain-behavior' paradigm originated under trait psychology and evolved further under behaviorist psychology. From trait psychology came the notions of characterizing characteristics of persons in terms of numbers on a measurement scale, and taking as evidence for these numbers, counts of keyed behaviors in samples from a domain of relevant settings (such as test items). The following quotation reflects how this perspective came to be applied to the development and practice of educational assessment:

The educational process consists of providing a series of environments that permit the student to learn new behaviors or modify or eliminate existing behaviors and to practice these behaviors to the point that he displays them at some reasonably satisfactory level of competence and regularity under appropriate circumstances. The statement of objectives becomes the description of behaviors that the student is expected to display with some regularity. The evaluation of the success of instruction and of the student's learning becomes a matter of placing the student in a sample of situations in which the different learned behaviors may appropriately occur and noting the frequency and accuracy with which they do occur (Krathwohl and Payne, 1971, p. 17-18).

Under the domain-behavior approach, the specification of an assessment describes a collection of task contexts as seen from the assessor's point of view, and provides a system for classifying the responses students might make. Potential responses in some contexts, such as multiple-choice items, are unambiguously right or wrong; in others, counts or instances of behaviors of certain types, the distinction of which may require expert judgment, are recorded. Behavior observed in a sample of tasks constitutes direct evidence for expected behavior in the domain as a whole, which in turn constitutes an operational definition of competence. The primary inferential task of standard test theory is to characterize the weight of evidence that samples of tasks provide about students' domain proficiencies. The processes by which students *acquire* competence are of interest, of course, to students, teachers, and researchers alike, but for the most part, these questions lie outside the universe of discourse associated with the domain-behavior paradigm of assessment (Stake, 1991).

In contrast, the acquisition of competence plays a central role in contemporary cognitive and educational psychology. The following quotation reflects the cognitive/developmental perspective as it relates to educational assessment:

Essential characteristics of proficient performance have been described in various domains and provide useful indices for assessment. We know that, at specific stages of learning, there exist different integrations of knowledge, different forms of skill, differences in access to knowledge, and differences in the efficiency of performance. These stages can define criteria for test design. We can now propose a set of candidate dimensions along which subject-matter competence can be assessed. As competence in a subject-matter grows, evidence of a knowledge base that

is increasingly coherent, principled, useful, and goal-oriented is displayed, and test items can be designed to capture such evidence (Glaser, 1991, p. 26 emphasis in original)

From the cognitive perspective, the specifications for an assessment describe contexts that can evoke evidence about students' competence as conceived at a higher level of abstraction, and provide judgmental guidelines for mapping from observed behavior to this inferred competence. This behavior provides evidence about competence so conceived, but not necessarily *direct* evidence. We may have to interpret this behavior in light of additional knowledge or supporting evidence about, for example, how the content or the context of a task interacts with the student; we may need to infer, or learn more about, the task as seen from the point of view of the student.

The ACTFL reading proficiency guidelines (Table 1) illustrate this point. Contrast the description of Intermediate readers' competence with texts 'about which the reader has personal interest or knowledge' with Advanced readers' competence with '...texts which treat unfamiliar topics and situations.' This distinction is fundamental to the underlying conception of developing language proficiency, but obviously a situation that is familiar to one student is unfamiliar to others. The evidential import of the same behavior in the same situation can differ radically for different students, and, as we shall explore further, affect what we infer about their capabilities from their behavior.

Probability-Based Inference

Probability isn't really about numbers; it's about the structure of reasoning.

Glenn Shafer (quoted in Pearl, 1988, pp. 44)

As the preceding section addressed what we want to reason about in educational assessment, this section concerns how we want to reason. It outlines the basic kinds of reasoning tasks we face, and reviews some tools from probability theory we can gainfully employ to this end, some hundreds of years old and others quite recent.

Kinds of Inference

Schum (1987) distinguishes among deductive, inductive, and abductive reasoning, all of which play essential and interlocking roles in educational assessment:

- *Deductive reasoning* flows from generals to particulars, within an established framework of relationships among variables—from causes to effects, from diseases to symptoms, from the way a crime is committed to the evidence likely to be found at the scene, from a student's knowledge and skills to observable behavior. Under a given state of affairs, what are the likely outcomes?
- *Inductive reasoning* flows in the opposite direction, also within an established framework of relationships—from effects to possible causes, from symptoms to